IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 November 2023, accepted 19 December 2023, date of publication 28 December 2023,
date of current version 10 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347410

RESEARCH ARTICLE

DMRNet Based Tuberculosis Screening With Cough Sound

WENLONG XU[1], HAIXIN YUAN[1], XIAOMIN LOU[2], YUANYUAN CHEN[2], AND FENG LIU[3]
[1]College of Information Engineering, China Jiliang University, Hangzhou, Zhejiang 310018, China
[2]Hangzhou Red Cross Hospital, Hangzhou, Zhejiang 310011, China
[3]School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, QLD 4702, Australia

Corresponding author: Wenlong Xu (wenlongxu@cjlu.edu.cn)

This work was supported in part by the Research and Development Projects of Zhejiang Province of China under Grant 2023C03107 and Grant 2020ZJZC02, and in part by the Hangzhou Social Development Research Project 202004A10.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Medical Ethics Committee of Red Cross Hospital Hangzhou in China under Approval No. 5[2023].

ABSTRACT Tuberculosis is the leading cause of death due to a single infection prior to the COVID-19 pandemic. Screening of tuberculosis patients in a large population is of paramount importance for disease treatment and control, especially with an economical, accurate, and easy-to-operate method. Based on cough sounds, we proposed DMRNet to distinguish between patients with tuberculosis, other respiratory diseases, and healthy individuals. DMRNet comprises four convolutional blocks and six identification blocks and incorporates dynamic convolution into the first three convolutional blocks to promote feature extraction. After the second and third dynamic convolutions, a polarized self-attention mechanism was added to reduce the information loss caused by the dimensionality reduction. Finally, a multihead self-attention layer is added to the fourth convolutional block and the last three identification blocks to enhance the aggregation of global information. Using a dataset with 1323 cough sound fragments, the results achieved the accuracy, sensitivity, and specificity of tuberculosis screening were 94.32%, 97.73%, and 99.43%, respectively. Compared with reported studies, the proposed model demonstrated better accuracy and reliability. Cough sound-based DMRNet analysis is a promising method for tuberculosis screening, especially in densely populated areas. Owing to its convenience, low equipment requirements, and low cost, it is expected to become an effective tool for community tuberculosis screening.

INDEX TERMS Convolutional neural networks, cough sounds, tuberculosis screening.

I. INTRODUCTION

According to the Global Tuberculosis Report 2022, there are 10.6 million new tuberculosis (TB) cases worldwide by 2021, with the number of deaths due to TB reaching 187,000 [1]. Before the COVID-19 pandemic, as an infectious disease [2], TB was the leading cause of death owing to a single infectious source, ranking higher than HIV. TB is caused by Mycobacterium tuberculosis, which spreads when an infected person excretes bacteria into the air, typically through coughing. Despite significant developments in screening, triage, and diagnostic methods for TB detection, large-scale screening remains challenging. Timely detection of TB is crucial for treatment and disease control.

Currently, the recommended methods for screening TB include chest radiography, interferon-gamma release assays (IGRA), tuberculin skin tests (PPD), and recombinant Mycobacterium TB fusion protein (EC). Chest radiography is the most commonly used method for TB screening [3]. The World Health Organization stipulates that chest radiography should be used when shunting and screening for active TB [4]. It uses X-rays to produce a two-dimensional image map of chest tissues, allowing for assessment of the internal structure of the chest and detection of diseases. This is the main method

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

used to clinically examine respiratory diseases and evaluate organ lesions. However, chest X-rays contain ionizing radiation, and should not be used in excess. The IGRA test, an in vitro immune diagnostic test, requires the collection of human peripheral whole blood for testing on specific laboratory instruments with a sensitivity and specificity of 88% and 63%, respectively [5]. However, the IGRA test is relatively expensive, requires specific laboratory conditions, and is limited by high-throughput testing, which results in cost difficulties for large-scale screening at the grassroots level. PPD, a skin test that uses a pure protein derivative as a common reaction source, is simple, convenient, and does not require specialized equipment or laboratories. Therefore, it is widely used as an auxiliary diagnostic tool in the clinical practice. However, this test often generates more false positives, particularly among individuals receiving the BCG vaccine [6], and repeated testing may also produce false positives. The EC test is a new type of skin test for Mycobacterium TB infection that is inexpensive and easy to perform. However, the results of this experiment were not as accurate as those of the IGRA trial were. Therefore, there is an urgent need for a convenient, non-invasive, low-cost, and highly accurate TB screening method.

Tracey et al. [7] developed a cough detection algorithm to monitor the recovery of TB patients. They extracted Mel-frequency cepstral coefficients from the audio signals of 10 subjects and used a combination of artificial neural network (ANN) and support vector machine (SVM) classifiers to detect coughs, with an overall sensitivity of 81%. Botha et al. [8] collected cough sound data from 17 patients with TB and 21 healthy subjects and then used logistic regression (LR) classifiers for classification, achieving an accuracy of 78%. In addition to audio recordings, five objective clinical measurements (MUAC, Temperature, BMI, Anaemic conjunctivae, Heart rate) were collected for each patient. When these additional 5 measurements were added, the accuracy improved to 82%. Pahar et al. [9] collected cough sounds from 16 TB patients and 35 patients with other respiratory diseases and used five machine learning classifiers (LR, SVM, k-nearest neighbor (kNN), multilayer perceptron, and convolutional neural networks (CNN)) to automatically recognize the two. LR was used to achieve the best performance, and sequential forward selection was added as a feature-selection method. The system achieved a sensitivity of 93% and a specificity of 95%. Pahar et al. [10] subsequently collected cough sound data from 47 TB patients, 229 COVID-19 patients, and 1498 healthy subjects, and used the pre-trained ResNet50 to distinguish between patients with TB and patients with COVID-19, with the highest F1 score of 0.9259. To distinguish patients with TB from those with COVID-19 and healthy subjects, the highest F1 score was 0.8631. These studies indicate that TB screening using audio analysis is practical. Owing to the relatively limited number of tuberculosis cases, the effectiveness of this method needs to be studied on a large scale. Although this paper has a larger sample size in the number of cases and cough sounds compared to reference [7], [8], [9], [10], especially in TB patients' number, deep leaning is a data-driven model [11], and an increase in sample data size will definitely have a direct impact on improving accuracy. The above studies produced results using existing deep learning models or machine learning methods, and no more advanced models have been proposed for classification experiments.

Therefore, based on cough sounds, we propose a classification model, DMRNet, to distinguish between patients with TB, other respiratory diseases (RDs), and healthy subjects. Our goal was to improve the accuracy of screening and to take necessary measures to reduce the spread of the virus. To verify the effectiveness of the method, comparative experiments were conducted, and the experimental results showed that the method achieved good performance in terms of accuracy, sensitivity, and specificity compared to current advanced deep learning models.

II. DATA AND METHODS

A. DATA DESCRIPTION

The cough sound data were recorded in a quiet ward using a smartphone. The sound files were saved in format. During the recording, the mask was removed if the patient wore it. The microphone was placed 30-40 cm away from the participant's mouth, at 45° horizontally upward. Participants were asked to cough actively. The standard cough is to take in full air and cough out forcefully. The dataset consists of records from 345 subjects, including 141 patients with TB, 52 with other respiratory diseases, and 152 healthy subjects. Because of the possibility of other non-cough sound audio recordings during the collection process, cough sound segments were extracted from the audio recordings at a sampling frequency of 44,100 Hz. Owing to the limited number of cough sounds in other respiratory diseases, time-shifting, pitch-shifting, and time-stretching have been used for data augmentation to achieve class balance. The specific distributions of the data are listed in Table 1.

B. MEL SPECTROGRAM

Because cough sounds have more energy at lower frequencies, the Mel spectrogram can provide more information at lower frequencies. We converted the cough sound data into Mel spectrograms. The Mel spectrogram contains the short-time Fourier transform of each frame of the spectrum, ranging from a linear frequency scale to a logarithmic Mel scale, and it then passes through a filter bank to obtain the feature vectors. By converting the processed cough sound data into Mel spectrograms, we trained the convolutional neural network to screen for TB.

C. PROPOSED METHOD

A flowchart of the proposed TB screening method is shown in Fig. 1. The cough sound fragments are input and then converted into Mel spectrograms and finally into a convolutional neural network for classification. Owing to the success of the

TABLE 1. Data distribution.

Disease Group	Number of Recordings	Age Range (Gender Known)			Total Number of Coughs
		18-40 (male:female)	41-65 (male:female)	65-100 (male:female)	
Healthy	152	15:8	36:30	32:31	441
TB	141	17:20	45:15	29:15	441
RDs	52	3:2	17:8	16:6	441

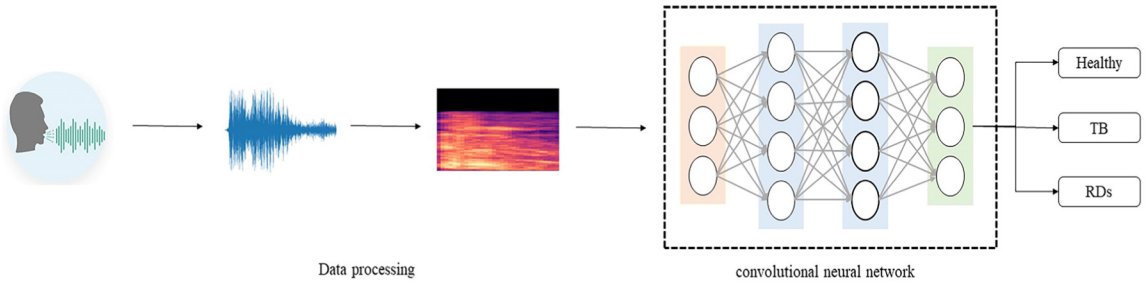

FIGURE 1. The flowchart of the proposed method.

ResNet structure in image recognition and computer vision tasks, and the fact that ResNet50 is considered one of the models with the highest accuracy in diagnosing COVID-19 from audio samples [10], [12], we proposed a similar deep convolutional neural network to classify diseases through audio by converting cough audio signals into image signals. The DMRNet proposed in this study was inspired by the CIdeR [13] network, and its structure was composed of appropriate ResNet blocks. DMRNet achieves good results after deleting and recombining based on ResNet50. Thus, we concluded that a deeper network does not have a better effect. DMRNet retains only four convolutional blocks and six identification blocks. The convolutional blocks first reduce the dimensionality of the feature image through convolution of 1×1, then perform a dynamic convolution operation, and finally restore the dimensionality through convolution of 1×1, followed by batch normalization and Gaussian error Linear Unit (GeLU) layers. In the other branch, a convolutional network of 1×1 is used to reduce the dimensionality of the maxpool output. The main branch of the recognition block was the same as that of the convolutional block. However, the other branch did not undergo dimensionality reduction through the convolutional network and directly added the input to the final 1×1 convolutional output. Dynamic convolution is added to the first three convolutional blocks for more flexible feature extraction and a more stable training process. Adding the PSA mechanism after the second and third dynamic convolutions can reduce the information loss caused by the dimensionality reduction. Adding the MHSA layer to the fourth convolutional block and the last three identification blocks effectively improves the overall learning performance of the model. The entire network used the GeLU function, thereby reducing the problem of gradient disappearance during the training process and accelerating the convergence of the model. The overall network architecture of the DMRNet is illustrated in Fig. 2.

1) DYNAMIC CONVOLUTION MODULE

We added dynamic convolution operations to the first three convolutional blocks, which did not increase the depth or width of the network. It can dynamically generate convolutional kernel coefficients according to different inputs, and then adaptively aggregate the kernels for better feature extraction and representation. The framework is illustrated in Fig. 3.

The traditional perceptron is denoted as $y = g\left(W^T x + b\right)$, where W and b represent the weight matrix and bias vector, respectively, and g represents the activation function. We defined the dynamic perceptron by aggregating the K linear functions $\left\{\widetilde{W}_k^{\mathrm{T}} x + \widetilde{b}_k\right\}$:

$$y = g\left(\widetilde{W}^T(x)\, x + \widetilde{b}(x)\right)$$
$$\widetilde{W}(x) = \sum_{k=1}^{K} \pi_k(x)\, \widetilde{W}_k, \widetilde{b}(x) = \sum_{k=1}^{K} \pi_k(x)\, \widetilde{b}_k$$
$$s.t.\ 0 \le \pi_k(x) \le 1, \ \sum_{k=1}^{K} \pi_k(x) = 1 \tag{1}$$

where π_k is the attention weight of the k^{th} linear function $\widetilde{W}_k^T x + \widetilde{b}_k$ and the aggregation weights $\widetilde{W}(x)$ and bias $\widetilde{b}(x)$ are input functions that share the same attention. The attention weight $\{\pi_k(x)\}$ changes as input x changes. For a given input, they represent the optimal aggregate $\left\{\widetilde{W}_k^T x + \widetilde{b}_k\right\}$ of the linear model, which is a nonlinear function. Therefore, the representation ability of the dynamic perceptron was better than that of the traditional perceptron.

The number of output channels of the dynamic perceptrons was the same as that of the traditional perceptrons; however, the model scale of the dynamic perceptrons was larger.

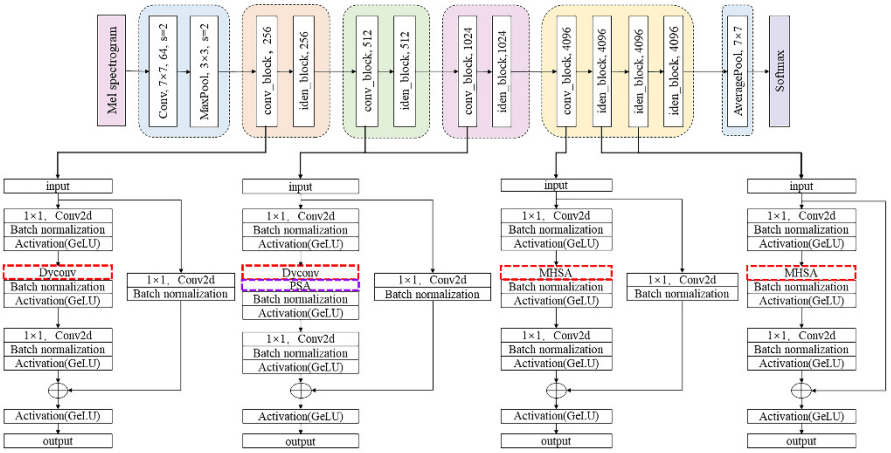

FIGURE 2. Overall framework diagram of DMRNet model.

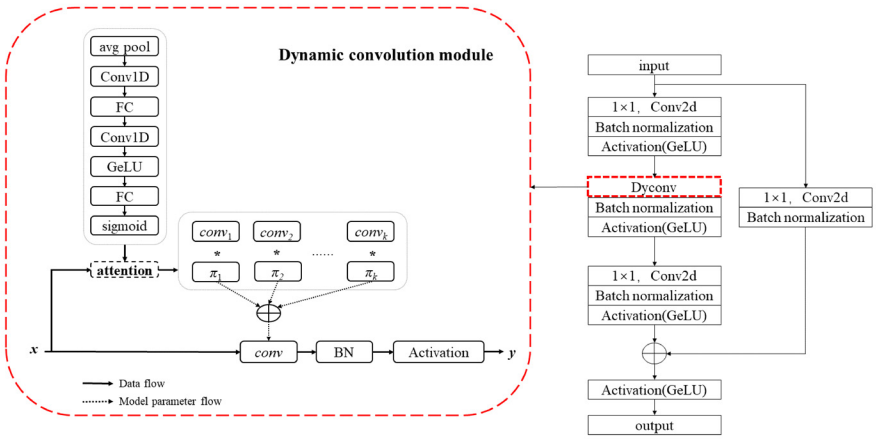

FIGURE 3. Dynamic convolution module.

Therefore, the dynamic perceptron was used for two additional calculations: (a) calculating the attention weights $\{\pi_k(x)\}$, and (b) the aggregation parameters based on the attention $\sum_k \pi_k \widetilde{W}_k$ and $\sum_k \pi_k \widetilde{b}_k$. The cost of the additional calculation should be significantly lower than taht of $\widetilde{W}^T x + \widetilde{b}$. Mathematically, the calculation constraints are expressed as follows:

$$O\left(\widetilde{W}^T x + \widetilde{b}\right) \gg O\left(\sum \pi_k \widetilde{W}_k\right) + O\left(\sum \pi_k \widetilde{b}_k\right) + O(\pi(x)) \quad (2)$$

where $O(\cdot)$ represented the cost of measurement calculation.

Dynamic convolution can effectively address these computational constraints. The principle of dynamic convolution is similar to that of dynamic perceptron. It has K convolutional kernels of the same size, aggregated using the attention weights $\{\pi_k\}$, and subsequently incorporated batch normalization and activation functions.

We improved Squeeze and Excitation (SE) [14] and ECAnet [15] to calculate the kernel attention $\{\pi_k(x)\}$. First, global average pooling was used to compress global spatial information. Subsequently, the attention weights of the K convolution kernels were generated using two one-dimensional convolutions, two fully connected layers, and Softmax. The second one-dimensional convolution and the second fully connected layer also contain a GeLU [16] function. Unlike SENet, dynamic convolution focuses on the convolutional kernels.

2) PSA MODULE

Most pixel-level regression deep convolutional neural networks adopt a common backbone network such as ResNet [17], which is frequently used in classification and regression tasks. However, these backbone networks can reduce spatial resolution while improving channel resolution for robustness and computational efficiency, resulting in

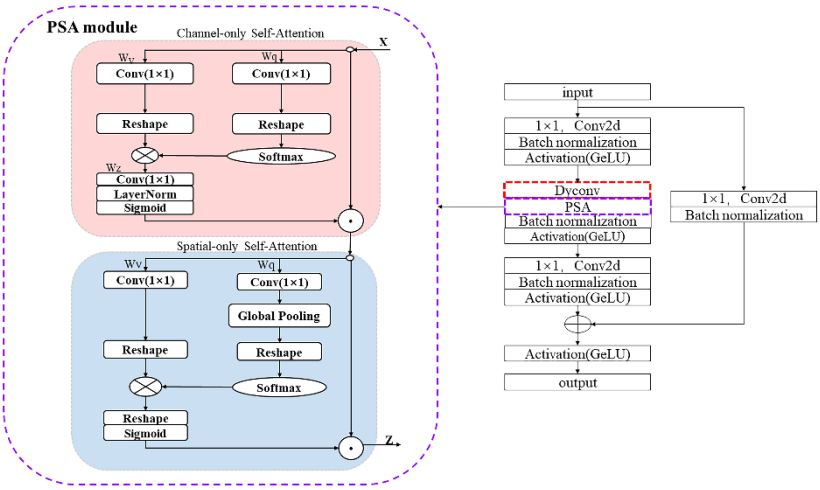

FIGURE 4. PSA module.

low-resolution features. Resolution loss makes pixel-level regression challenging, particularly in cases with a high degree of nonlinearity at object edges and body parts, making it difficult to encode information with low-resolution features [18], [19], [20].

To effectively utilize high-resolution information, we implemented "polarization filtering" in the attention calculation. A self attention block is operated on the input tensor $\mathbf{X}$ to highlight or suppress the features. We introduced a PSA mechanism in the convolutional blocks. This mechanism completely folded the features in one direction for filtering operations, while maintaining a high resolution in orthogonal directions. After normalizing the bottleneck tensor using the Softmax function, we used the Sigmoid function for tone mapping to expand the dynamic range of attention. We operationalized the PSA mechanism into a PSA module, as shown in Fig. 4.

The channel-only branch $A^{ch}(\mathbf{X}) \in \Re^{C\times 1\times 1}$ denoted as follows:

$$\mathbf{A}^{ch}(\mathbf{X}) = F_{SG}\left[\mathbf{W}_{z|\theta_1}\left(\left(\sigma_1\left(\mathbf{W}_v(\mathbf{X})\right)\right) \times F_{SM}\left(\sigma_2\left(\mathbf{W}_q(\mathbf{X})\right)\right)\right)\right] \quad (3)$$

where $\mathbf{W}_q$, $\mathbf{W}_v$, and $\mathbf{W}_z$ are all convolutional layers of 1×1, σ_1, and σ_2 are tensor-reshaping operators, and $F_{SG}(\cdot)$ and $F_{SM}(\cdot)$ are the Sigmoid and Softmax operators, respectively. The output of the channel-only branch is $\mathbf{Z}^{ch} = \mathbf{A}^{ch}(\mathbf{X}) \odot^{ch} \mathbf{X}$, where $\odot^{ch}$ is a channel-by-channel multiplication operator.

The spatial-only branch $A^{sp}(\mathbf{X}) \in \Re^{1\times H\times W}$ denoted as follows:

$$\mathbf{A}^{sp}(\mathbf{X}) = F_{SG}\left[\sigma_3\left(F_{SM}\left(\sigma_1\left(F_{GP}\left(\mathbf{W}_q(\mathbf{X})\right)\right)\right) \times \sigma_2\left(\mathbf{W}_v(\mathbf{X})\right)\right)\right] \quad (4)$$

where $\mathbf{W}_q$ and $\mathbf{W}_v$ are the convolutional layers of 1×1, σ_1, σ_2, and σ_3 are the three tensor-reshaping operators, $F_{SG}(\cdot)$ and $F_{SM}(\cdot)$ are the Sigmoid and Softmax operators respectively, $F_{GP}(\cdot)$ is the global pooling operator, and the output of the spatial-only branch is $\mathbf{Z}^{sp} = \mathbf{A}^{sp}(\mathbf{X}) \odot^{sp} \mathbf{X}$, where $\odot^{sp}$ is a spatial-by-spatial multiplication operator. The PSA module combines the outputs of the two branches in parallel or series.

In contrast to other attention modules, the PSA module maintains a high level of attention resolution in both channel and spatial dimensions. In addition, we combined Softmax reweighting with squeezing excitation in channel-only attention, using Softmax as a nonlinear activation at the bottleneck tensor. The number of channels follows a squeezing excitation mode, which is advantageous for both GC [21] and SE blocks, achieving higher-resolution squeezing and excitation. The spatial-only branch attention maintains the complete spatial resolution and internally preserves the learnable parameters in W_q and W_v for nonlinear Softmax reweighting, thus offering a more robust structure than the existing modules.

The Softmax-Sigmoid combination was used for both the channel-only and spatial-only PSA branches. Using the Softmax-Sigmoid combination as a probability distribution function, multimodal Gaussian and segmented binomial maps can be approximated using linear transformations. Therefore, the nonlinearity can fully exploit the high-resolution information retained in the PSA module branch.

3) MHSA MODULE

Although CNN focus on aggregating local information, modelling long-range dependencies in visual tasks is typically necessary. It is desirable to stack more convolutional layers [17]. Although more stacked layers improve the performance [22], the mechanism for modelling global dependencies is a powerful and scalable solution that does not require stacking many layers.

An efficient way to use self-attention [23] in the model is to replace spatial convolution with the MHSA layer pro-

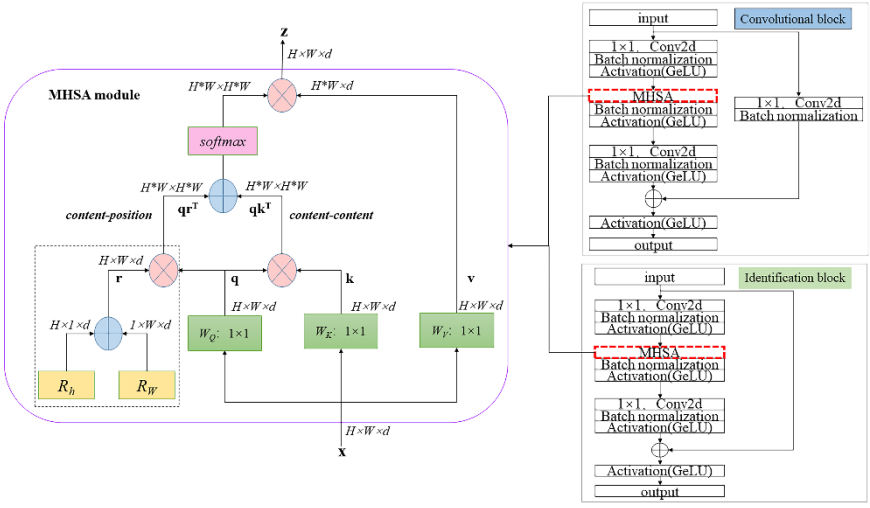

FIGURE 5. MHSA module.

posed in the transformer. We referred to models such as SASA [24], AACN [25], SANET [26], Axial-ASA [27], and Vision Transformer (VIT) models [28], which use different self-attentions to replace spatial convolution in ResNet bottleneck blocks [17]. We used convolution to learn abstract and low-resolution feature maps from images, and used global attention to process the information in the feature maps captured by convolution. Non-local (NL) nets [29] established a connection between the transformer and the non-local mean algorithm [30]. They inserted NL blocks into the last two blocks of the ResNet to improve the performance of the target task. Similar to NL Net [21], [29], a hybrid design of convolution and global self-attention was used in our model. However, in our model, we use MHSA layers instead of convolutions in the last four blocks to achieve global self-attention in the feature map. The framework is illustrated in Fig. 5.

In the DMRNet, transformer-based architectures are used for location awareness [23]. This model implements the 2D relative position self-attention from [24], [25]. This is very similar to the multihead self-attention in the transformer, but the difference is that MSHA treats position encoding as spatial attention, embedding two learnable vectors as spatial attention in both the horizontal and vertical dimensions. The added and fused spatial vectors are then multiplied by q to obtain the content position, and the content position and content are multiplied to obtain spatially sensitive similarity features, allowing MHSA to focus on more suitable regions and easier convergence.

4) GELU FUNCTION

Because of its advantages such as nonlinearity, differentiability, and smoothness, we used the GELU [16] activation function.

The neuron input x is multiplied by m$\sim$ Bernoulli $(\Phi(x))$, $\Phi(x) = P(X \leq x)$, where $X \sim \mathbf{N}(0, 1)$ is the cumulative distribution function of a standard normal distribution. A new nonlinearity arises when obtaining deterministic decisions from neural networks. Nonlinearity is the expected transformation of the random regularizer on input x, that is, $\Phi(x) \times Ix + (1 - \Phi(x)) \times 0x = x\Phi(x)$. In short, x is scaled based on how much larger x is compared to other inputs. Because the cumulative distribution function of the Gaussian functions is usually calculated using an error function, the GELU is defined as

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{5}$$

Approximately

$$0.5x\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + 0.44715x^3\right)\right]\right) \tag{6}$$

III. EXPERIMENTS

A. EXPERIMENTAL SETUP

The dataset used in this experiment is divided into three parts: 60% for training, 20% for validation, and 20% for testing. The optimizer was set to Adamax, with an initial learning rate of 0.001. The batch size was set to 32 and the number of epochs was set to 300. If the loss function value of the validation set did not change after 30 epochs, the learning rate was decreased to half of the original value. If the accuracy of the validation set did not improve after 50 epochs, then the iteration was stopped.

The operating system used in this article was Ubuntu 18.04.6 LTS, the CPU used Intel (R) Core (TM) i7-1065G7 CPU@1.30GHz 1.50GHz, and the GPU used GeForce RTX 3090 (24.0GB). The language environment was Python 3.6, and the framework was TensorFlow 2.6.2.

TABLE 2. Performance indicators of cough diagnosis classifier.

	Precision	Sensitivity	Specificity	F1-score	Accuracy
Overall	-	-	-	94.34%	94.32%
TB	98.85%	97.73%	99.43%	98.29%	-
Healthy	88.42%	95.45%	93.75%	91.80%	-
RDs	96.34%	89.77%	98.29%	92.94%	-

TABLE 3. Ablation experiments.

	Sensitivity			Specificity			Accuracy	F1-score
	TB	Healthy	RDs	TB	Healthy	RDs		
DMRNet	97.73%	95.45%	89.77%	99.43%	93.75%	98.29%	94.32%	94.34%
DMRNet-MHSA	96.59%	92.04%	85.23%	95.45%	92.04%	99.43%	91.29%	91.30%
DMRNet-MHSA-PSA(Spatial-only Self-Attention)	95.45%	94.32%	89.77%	100%	93.74%	96.02%	93.18%	93.23%
DMRNet-MHSA-PSA	96.59%	85.22%	93.18%	99.43%	94.88%	93.18%	91.67%	91.67%
DMRNet-MHSA-PSA-Dynamic	90.91%	88.63%	89.77%	97.15%	92.04%	95.45%	89.77%	89.81%

TABLE 4. Comparison of results with commonly used deep learning models.

Model		Sensitivity			Specificity			Accuracy	F1-score
		TB	Healthy	RDs	TB	Healthy	RDs		
VGG	16 layer	92.04%	73.86%	89.77%	95.45%	93.18%	89.20%	85.22%	85.23%
	19 layer	94.32%	71.59%	82.95%	96.59%	89.77%	88.07%	82.95%	82.95%
ResNet	18 layer	93.18%	88.64%	85.23%	98.29%	89.77%	95.45%	89.02%	89.10%
	34 layer	93.18%	94.32%	88.64%	98.86%	91.48%	97.72%	92.05%	92.05%
	50 layer	98.86%	84.09%	85.22%	90.91%	94.31%	98.86%	89.34%	89.35%
	101 layer	90.91%	87.55%	90.91%	98.85%	91.48%	94.31%	89.77%	89.77%
Xception		87.50%	85.22%	71.59%	98.29%	84.09%	89.77%	81.44%	81.59%
MobileNet		92.04%	82.95%	86.36%	97.72%	89.77%	93.18%	87.12%	87.19%
DenseNet121		89.78%	86.36%	78.41%	99.43%	85.89%	92.04%	84.85%	85.06%
GoogleNet		90.91%	70.45%	71.59%	97.16%	84.09%	85.22%	77.65%	77.78%
InceptionV3		86.36%	67.05%	84.09%	95.45%	89.20%	84.09%	79.17%	79.12%
DMRNet		97.73%	95.45%	89.77%	99.43%	93.75%	98.29%	94.32%	94.34%

B. RESULTS

The evaluation indicators used were sensitivity, specificity, accuracy, and the F1 score. The F1 score is the harmonic mean of the accuracy and recall, with a maximum value of 1 and minimum value of 0. The higher the F1 score, the better is the performance of our model. Table 2 shows the classification performance of DMRNet, with an accuracies and F1 score of 94.32% and 94.34%, respectively.

To evaluate the effectiveness of each module of the model, ablation experiments were conducted; and the results are presented in Table 3. When MHSA was replaced with an 3×3 convolution, the accuracy decreased by 3.03% compared to that of DMRNet. On this basis, by removing the spatial self-attention mechanism in PSA, the accuracy was improved by 1.89%, but decreased by 1.14% compared with DMRNet. If we continue to remove the entire PSA module, the accuracy would be only 91.67%. Finally, by replacing the dynamic convolution with an 3×3 convolution, the accuracy was reduced by 4.55% compared with that of DMRNet. Addition of various modules yielded the best results.

Table 4 compares the proposed model with existing common deep-learning models. The table indicates that a deeper network may not yield a better classification performance. As can be seen from Table 4, among these common deep learning models, Google Net had the worst classification performance, with an accuracy rate of only 77.65%, which is 16.67% lower than that of the proposed model. ResNet34 had the best classification performance, achieving an accuracy 92.05% and 2.27% lower than that of the proposed model. The models with the lowest sensitivity and specificity for TB screening were InceptionV3 and ResNet50, which were 86.36% and 90.91%, respectively, whereas the models with the highest sensitivity and specificity were ResNet50 and DenseNet121, which were 98.86% and 99.43%, respectively. ResNet50 had a sensitivity 1.13% higher than that of DMRNet, whereas DenseNet121 and DMRNet had the same specificity, but the overall classification effect of both models was not as good as that of DMRNet.

As shown in Table 5, we compared the model with existing studies. Sensitivity, specificity, accuracy, and F1 scores were

TABLE 5. Comparison results with existing methods.

	Sensitivity			Specificity			Accuracy	F1-score
	TB	Healthy	RDs	TB	Healthy	RDs		
Pahar et al. [10]	92.05%	80.68%	76.14%	98.29%	85.23%	90.91%	82.95%	83.10%
Rahman et al.[31]	95.45%	84.09%	81.82%	99.43%	89.77%	91.47%	87.12%	87.21%
Son et al.[32]	92.04%	87.50%	78.41%	98.29%	88.07%	92.61%	85.98%	86.05%
Imran et al.[33]	96.59%	87.50%	79.55%	99.43%	88.07%	94.32%	87.87%	87.88%
DMRNet	97.73%	95.45%	89.77%	99.43%	93.75%	98.29%	94.32%	94.34%

used as metrics to compare the classification performance of the different models. The results in the table indicate that the proposed model exhibited the best classification performance.

IV. DISCUSSION

Because of the powerful representation ability of deep neural networks [34], they have achieved remarkable results on many problems; therefore, we chose ResNet as the basic framework. The core architecture consisted of four convolutional blocks and six identification blocks stacked in the order of one convolutional block and one identification block, respectively, with the remaining two blocks stacked at the end. Adding dynamic convolution operations to the first three convolutional blocks allows for more flexible feature extraction and improved accuracy. Dynamic convolution dynamically aggregates multiple convolution kernels according to the degree of attention of the attention module for each input. Compared with static convolutional kernels, this approach significantly improves representational capacity. In addition, the attention computing module, composed of SE and ECA, makes the training process more stable.

After the second and third dynamic convolutions, the PSA module was added to better handle short- and long-distance dependencies. Both the PSA channel-only and spatial-only branches used a Softmax-Sigmoid combination. The combination of Softmax-Sigmoid as a probability distribution function minimized the potential loss of high-resolution information as much as possible. Similar to CBAM, PSA makes series or parallel connections only between the channel and spatial branches [35]. However, structures with dual attention, such as CBAM, mainly obtain attention weights through fully connected and convolutional layers; therefore, they may not be effective for mining information. However, in this study, we used self-attention to obtain attention weights using the modelling ability of the self-attention structures. In addition, we implemented a feature dimension reduction operation to achieve highly effective long-distance modelling.

In this study, MHSA layers were used instead of the convolution operations in the last four blocks of the model. These layers achieve global self-attention on the feature map. Global self-attention was used to process and aggregate the information in the feature maps captured by the convolution. This hybrid structure combines the advantages of CNN and self-attention, utilizing convolution for spatial downsampling and focusing self-attention on smaller resolution. The accuracy of the final model classification was significantly improved. A common technique involves using a location-coding system structure based on a transformer to enable location-aware attention operations. Relative position encoding facilitates attention to focus on content information and the relative distance between the features at different locations. This effectively associates the cross-object information with positional awareness. Finally, when the two branches in the PSA were connected in series, the model achieved the best performance.

In summary, the ablation experiments emphasized the importance of each module in the DMRNet model. The combination of these modules enabled DMRNet to achieve excellent performance across all evaluation indicators, demonstrating its effectiveness in using cough sounds to distinguish between TB patients, healthy subjects, and those with other respiratory diseases. From the results in Tables 4 and 5, DMRNet learned good features. This is evident because our method achieves the highest accuracy compared with the other methods.

The dataset was collected from hospital wards in relatively quiet environments. The intercepted cough audio had a high signal to noise ratio. It is possible that classification performance decreases when the data are sampled in a community with strong background noise. Although the volume of the dataset is relatively larger than that reported in previous studies, data accumulation is still the key to improving performance.

V. CONCLUSION

In this study, we proposed a tuberculosis-screening model based on a CNN that provides better performance in terms of accuracy and F1 score. The model consisted of two main parts. The first step was to convert the input cough sounds into a Mel spectrograms. The spectrogram itself fully contains the spectral information of the audio signal, which has not undergone any processing. The Mel spectrogram is based on the spectrogram and is subjected to Mel filtering, which makes the features more consistent with human recognition. The second part uses the deep convolutional model DMRNet to extract and classify features. We used convolutional neural networks to extract deep features of cough sound signals, instead of traditional methods, and achieved an accuracy of 94.32% in tuberculosis screening. In future, we plan to establish a larger dataset to improve the performance of the model.

Using cough sounds as a method of TB screening can provide a cost-effective solution for large-scale population

screening. When this technique is integrated into a user's daily activities, non-invasive data collection can be a more effective and user-friendly tool for individuals. Moreover, intelligent classification capability holds significant importance for hospitals and healthcare systems as it enables the objective analysis of users' symptoms and effectively reduces the burden on medical centers.

REFERENCES

[1] *Global Tuberculosis Report 2022*, World Health Org., Geneva, Switzerland, Oct. 2022, pp. 1–68. [Online]. Available: https://www.who.int/teams/global-tuberculosis-programme/tb-reports/globaltuberculosis-report-2022

[2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA, Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, Jan. 2019, doi: 10.3322/caac.21551.

[3] J. Devasia, H. Goswami, S. Lakshminarayanan, M. Rajaram, and S. Adithan, "Deep learning classification of active tuberculosis lung zones wise manifestations using chest X-rays: A multi label approach," *Sci. Rep.*, vol. 13, no. 1, p. 887, Jan. 2023, doi: 10.1038/s41598-023-28079-0.

[4] S. D. Boon and C. Miller, "Module 2: Screening-systematic screening for tuberculosis disease," in *WHO Consolidated Guidelines on Tuberculosis*. World Health Organization, 2021.

[5] C. M. Rumende, E. J. Hadi, G. Tanjung, I. N. Saputri, and R. Sasongko, "The benefit of interferon-gamma release assay for diagnosis of extrapulmonary tuberculosis," *Acta Med Indones*, vol. 50, no. 2, pp. 43–138, 2018.

[6] T. Wolf, U. Goetsch, G. Oremek, M. Bickel, P. Khaykin, A. Haberl, O. Bellinger, R. Gottschalk, H. R. Brodt, and C. Stephan, "Tuberculosis skin test, but not interferon-γ-releasing assays is affected by BCG vaccination in HIV patients," *J. Infection*, vol. 66, no. 4, pp. 376–380, Apr. 2013, doi: 10.1016/j.jinf.2012.11.004.

[7] B. H. Tracey, G. Comina, S. Larson, M. Bravard, J. W. López, and R. H. Gilman, "Cough detection algorithm for monitoring patient recovery from pulmonary tuberculosis," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Boston, MA, USA, Aug. 2011, pp. 6017–6020.

[8] G. Botha, G. Theron, R. Warren, M. Klopper, K. Dheda, P. Van Helden, and T. Niesler, "Detection of tuberculosis by automatic cough sound analysis," *Physiol. Meas.*, vol. 39, no. 4, Apr. 2018, Art. no. 045005, doi: 10.1088/1361-6579/aab6d0.

[9] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, and T. Niesler, "Automatic cough classification for tuberculosis screening in a real-world environment," *Physiol. Meas.*, vol. 42, no. 10, Oct. 2021, Art. no. 105014, doi: 10.1088/1361-6579/ac2fb8.

[10] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, A. Diacon, and T. Niesler, "Automatic tuberculosis and COVID-19 cough classification using deep learning," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET)*, Prague, Czech Republic, Jul. 2022, pp. 1–9.

[11] Y. Wu, S. Zhao, Z. Xing, Z. Wei, Y. Li, and Y. Li, "Detection of foreign objects intrusion into transmission lines using diverse generation model," *IEEE Trans. Power Del.*, vol. 38, no. 5, pp. 3551–3560, Oct. 2023, doi: 10.1109/tpwrd.2023.3279891.

[12] M. Pahar, M. Klopper, R. Warren, and T. Niesler, "COVID-19 cough classification using machine learning and global smartphone recordings," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104572, doi: 10.1016/j.compbiomed.2021.104572.

[13] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: A pilot study," *BMJ Innov.*, vol. 7, no. 2, pp. 356–362, Apr. 2021, doi: 10.1136/bmjinnov-2021-000668.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[15] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11531–11539.

[16] A. Nguyen, K. Pham, D. Ngo, T. Ngo, and L. Pham, "An analysis of state-of-the-art activation functions for supervised deep neural network," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Ho Chi Minh City, Vietnam, Aug. 2021, pp. 215–220.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[19] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 5686–5696.

[20] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021, doi: 10.1109/TPAMI.2020.2983686.

[21] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Seoul, South Korea, Oct. 2019, pp. 1971–1980.

[22] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 2736–2746.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[24] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," in *Proc. NIPS*, vol. 32, 2019.

[25] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3285–3294.

[26] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 10073–10082.

[27] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. ECCV*, 2020, pp. 108–126.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, Vienna, Austria, May 2021.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7794–7803.

[30] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. CVPR*, San Diego, CA, USA, 2005, pp. 60–65.

[31] T. Rahman, N. Ibtehaz, A. Khandakar, M. S. A. Hossain, Y. M. S. Mekki, M. Ezeddin, E. H. Bhuiyan, M. A. Ayari, A. Tahir, Y. Qiblawey, S. Mahmud, S. M. Zughaier, T. Abbas, S. Al-Maadeed, and M. E. H. Chowdhury, "QUCoughScope: An intelligent application to detect COVID-19 patients using cough and breath sounds," *Diagnostics*, vol. 12, no. 4, p. 920, Apr. 2022, doi: 10.3390/diagnostics12040920.

[32] M.-J. Son and S.-P. Lee, "COVID-19 diagnosis from crowdsourced cough sound data," *Appl. Sci.*, vol. 12, no. 4, p. 1795, Feb. 2022, doi: 10.3390/app12041795.

[33] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informat. Med. Unlocked*, vol. 20, 2020, Art. no. 100378, doi: 10.1016/j.imu.2020.100378.

[34] Z. Xing, S. Zhao, W. Guo, F. Meng, X. Guo, S. Wang, and H. He, "Coal resources under carbon peak: Segmentation of massive laser point clouds for coal mining in underground dusty environments using integrated graph deep learning model," *Energy*, vol. 285, Dec. 2023, Art. no. 128771, doi: 10.1016/j.energy.2023.128771.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. ECCV*, Munich, Germany, 2018, pp. 3–19.